



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2019

Transport mode detection based on mobile phone network data: A systematic review

Huang, Haosheng ; Cheng, Yi ; Weibel, Robert

Abstract: The rapid development in telecommunication networks is producing a huge amount of information regarding how people (with their mobile devices) move and behave over space and time. While GPS data, typically collected by smartphone apps, are restricted to rather small samples of the population, mobile phone network data, routinely collected by mobile network operators, potentially allow to analyze travel behaviors and social interaction of the whole population, with full temporal (e.g., longitudinal) coverage at a comparatively low cost. Therefore, recent years have seen an increasing interest in using such data for human mobility studies. However, due to their noisy and temporally infrequent/irregular nature, extracting mobility information such as transport modes from these data is particularly challenging. This paper provides an in-depth, systematic review of transport mode detection based on mobile phone network data. The results of the review show that existing studies tend to identify easy-to-detect modes (e.g., train or metro), or aggregate fine-grained modes into more general groups (e.g., public versus private transport). Rule-based methods making use of geographic data were often developed. More importantly, due to the lack of ground truth data, evaluation of the proposed methods was seldom done and reported. Finally, we identify a list of research gaps currently being faced in this field, particularly regarding robust and real-time data cleaning and mode detection methods, “benchmark” datasets and metrics allowing the comparison of different mode detection methods, as well as privacy and bias issues.

DOI: <https://doi.org/10.1016/j.trc.2019.02.008>

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-177487>

Journal Article

Accepted Version



The following work is licensed under a Creative Commons: Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.

Originally published at:

Huang, Haosheng; Cheng, Yi; Weibel, Robert (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, 101:297-312.

DOI: <https://doi.org/10.1016/j.trc.2019.02.008>

Transport Mode Detection Based on Mobile Phone Network Data: A Systematic Review

Haosheng Huang¹, Yi Cheng², Robert Weibel¹

¹GIScience Center, Department of Geography, University of Zurich, Zurich, Switzerland

²Institute of Geographic Information Engineering, Information Engineering University, Zhengzhou, China

Abstract

The rapid development in telecommunication networks is producing a huge amount of information regarding how people (with their mobile devices) move and behave over space and time. While GPS data, typically collected by smartphone apps, are restricted to rather small samples of the population, mobile phone network data, routinely collected by mobile network operators, potentially allow to analyze travel behaviors and social interaction of the whole population, with full temporal (i.e. longitudinal) coverage at a comparatively low cost. Therefore, recent years have seen an increasing interest in using such data for human mobility studies. However, due to their noisy and temporally infrequent/irregular nature, extracting mobility information such as transport modes from these data is particularly challenging. This paper provides an in-depth, systematic review of transport mode detection based on mobile phone network data. The results of the review show that existing studies tend to identify easy-to-detect modes (e.g., train or metro), or aggregate fine-grained modes into more general groups (e.g., public versus private transport). Rule-based methods making use of geographic data were often developed. More importantly, due to the lack of ground truth data, evaluation of the proposed methods was seldom done and reported. Finally, we identify a list of research gaps currently being faced in this field, particularly regarding robust and real-time data cleaning and mode detection methods, “benchmark” datasets, and metrics allowing the comparison of different mode detection methods, as well as privacy and bias issues.

Keywords: Mobile phone network data, systematic review, transport mode detection, human mobility, call detail records

1. Introduction

Understanding travel behaviors, travel demands, and the impact of transportation infrastructure has on people is central to transportation science. A widely used method to collect this type of data is through travel surveys or diaries, where users report why and how they travel to their destinations (Prelipean et al., 2017). Traditionally, this was done via face-to-face interviews, mail-out/mail-back with paper diaries, phone interviews, and Web forms (Shen and Stopher, 2014; Wu et al., 2016). These traditional approaches often lead to issues such as high respondent burden, significant implementation costs, poor data quality (e.g., low accuracy, lack of detail, under-reporting of trips) (Bricka and Bhat, 2006), and low response rates (Ogle et al., 2005). A key advantage of travel surveys is that they allow to probe on population mobility patterns, together with rich sociodemographic information, which also make them still one of the most reliable tools for transportation research. To overcome the issues of traditional travel surveys, transportation science has been trying to develop methods for automatically collecting travel diaries, and deriving important trip attributes such as start/end (where/when) of the trip, trip purposes, and travel modes.

The Global Positioning System (GPS), which provides accurate, time-stamped location data, is one of the most significant technologies to overcome the issues of traditional travel surveys since the late

1990s and early 2000s (Schönfelder et al., 2002; Wolf, 2000). Dedicated GPS loggers and more recently smartphones are used for GPS-based travel surveys. In recent years, a number of methods for processing GPS data have been developed to extract important trip characteristics for transportation research. Among others, many of the studies concentrated on inferring travel modes from GPS data, e.g., using rule-based methods, fuzzy logic, and machine learning (e.g., random forest, neural network) (Bantis and Haworth, 2017; Bohte and Maat, 2009; Dabiri and Heaslip, 2018; Prelipcean et al., 2017; Wu et al., 2016). GPS-based data collection methods present obvious advantages over traditional methods, as they require reduced effort of the respondents, provide greater spatial and temporal precision and detail, reduce labor and time costs of the researcher, and potentially allow longitudinal data collection over multiple days or even longer periods (particularly when GPS loggers are used) (Wu et al., 2016). However, GPS-based travel surveys require participants to carry a dedicated GPS logger or actively enable GPS tracking on their smartphone, which limits the scale and duration of this type of survey. Forgetting to take the GPS logger along or enable GPS tracking on smartphones will also result in incomplete data collection. Meanwhile, transferring data from smartphones or loggers to data centers is often needed, and might incur high Internet communication cost or require joint effort of researchers and respondents (they need to meet for the data transfer).

In addition to GPS data, recent research has also started exploring the use of other smartphone sensors for transport mode detection, such as Inertial Measurement Unit (IMU) sensors like accelerometers, magnetometers, and gyroscopes. For example, Feng and Timmermans (2013) investigated the benefits of combining GPS and accelerometer data for transport mode detection. Eftekhari and Ghatee (2016) developed methods to detect transport modes purely based on data from IMU sensors available on smartphones. However, these approaches often require users installing a special mobile application on their smartphone to collect the sensor data, which significantly limits the scale of this type of studies.

Other attempts to improve travel surveys in recent years saw the incorporation of data from automated fare collection systems, online geo-social media data, and particularly mobile telecommunication network data. The latter, such as call detail records (CDRs), handovers, and signaling data (Calabrese et al., 2014; Rojas et al., 2016), are often recorded automatically by telecommunication network operators for billing, management, and maintenance purposes, and therefore require no effort from the end users (they are not even aware the data are being recorded). These data often contain location information of cell phones, estimated by the cellular towers. According to International Telecommunication Union, the world coverage of mobile phone subscriptions has raised from 33.9% of the world population in 2005 to 103.5% in 2017 – 7.7 billion subscribers – corresponding to a penetration of 127.3% in developed countries and 98.7% in developing countries¹. Meanwhile, mobile phones are becoming one of the necessities in our daily life, and people very likely carry their phones the whole day. Therefore, these kinds of mobile phone network data potentially allow us to analyze travel behaviors of the whole population, with full temporal coverage at a comparatively low cost. However, these data tend to have lower spatial accuracy than GPS data, and are recorded infrequently and irregularly. Therefore, extracting mobility information such as transport modes from these data is particularly challenging, and requires proper processing methods.

The aims of this study are to provide an in-depth, state-of-the-art review on transport mode detection using mobile phone network data, assess existing data processing methods to show the advantages and disadvantages of each of them, and understand the research directions that need to be pursued. The methodology of a systematic review is employed, which has been used by many

¹ https://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2017/ITU_Key_2005-2017_ICT_data.xls

scientific disciplines (e.g., medicine and biology, social sciences, computer science) to minimize bias and ensure reproducibility of review studies (Budgen and Brereton, 2006; Moher et al., 2009; Petticrew and Roberts, 2006). We particularly focus on transport mode detection, mainly due to the following two reasons: 1) Understanding the modes of transport people take is key to travel behavior studies; 2) The process of transport mode detection often involves cleaning, segmentation, and inference, which are common to many mobility and urban planning applications that use mobile network data. Several recent review papers exist on the analysis of mobile phone network data, but they rather aim to provide an overview of the broad spectrum of their applications, including social network analysis, mobility analysis, event detection, and urban planning (Blondel et al., 2015; Calabrese et al., 2014; Naboulsi et al., 2016; Wang et al., 2018; Yuan and Raubal, 2016). Compared to these review papers, this systematic review paper focuses exclusively on transport mode detection using mobile phone network data. Particularly, we aim to answer the following questions:

- 1) Data: What kinds of mobile phone network data have been used for transport mode detection? What are the characteristics of these data?
- 2) Pre-processing: What are the existing pre-processing methods in transport mode detection based on mobile phone network data, particularly regarding noise removal and trip identification? What are their strengths and weaknesses?
- 3) Mode detection: What methods have been developed for transport mode detection? What are their strengths and weaknesses? What are the features mostly employed? To what extent are the mode detection methods evaluated?
- 4) Research gaps: What are the research gaps in transport mode detection using mobile phone network data?

It is important to note that similar terms have been also used in the literature to denote mobile phone network data, such as “mobile positioning data” (Ahas et al., 2010), “mobile phone data” (Steenbruggen et al., 2013; Wang and Chen, 2018), and “mobile network data” (Oliver et al., 2015).

The remainder of this paper is organized as follows. Section 2 briefly discusses characteristics of mobile phone network data. Section 3 then introduces the systematic review process. Results of the review study are presented in Section 4. Section 5 discusses the results, and answers the above questions. Finally, in Section 6 we summarize the findings, and suggest future research directions.

2. Mobile Phone Network Data

A typical cellular network consists of base transceiver stations (BTSs)², known as base stations or cell towers, each of which consists of one or more antennas and other equipment that facilitate wireless communication between a network and a mobile device (e.g., cellphones). Each BTS covers a defined area, known as a cell, which is the smallest spatial entity in the cellular network. A BTS is under control of a base station controller (BSC), which manages a set of BTSs, and is responsible for radio network management, call setup and handovers between its owning BTSs during call or data connection (e.g., when a user moves from one cell to another during a call). A BSC covers the cells of its owning BTSs, which together form a location area (LA)³. BSCs are grouped and controlled by Mobile Switching Centers (MSCs), which set up and release end-to-end connection (e.g., between mobile users, or from mobile users to other networks (e.g. telephone networks, other mobile networks, or internet networks)). They also handle mobility and hand-over requirements during the call or data connection, and take care of charging. An MSC often contains a visitor location register

² The term BTS, introduced originally in 2G networks (e.g., GSM), evolved with the development of next generations of mobile networks. Similar terms, such as Node B in 3G networks, or eNB (evolved Node B) in 4G networks, exist. However, this paper uses BTS as a generic term, implying all these similar terms.

³ Similarly, the term LA was introduced originally for 2G networks. 3G and 4G networks have similar terms, such as routing area and tracking area. This paper again uses LA to collectively denote these similar terms.

(VLR), which is a database recording the exact location of all mobile subscribers currently present in the service area of the MSC. See Figure 1 (left) for an illustrative example of a mobile phone network.

To assure the quality of communication services, cellular networks are constantly and frequently determining the location of mobile phones, even if they are simply on standby. The phone's location is often calculated by determining the location of the BTS it connects to. Therefore, the spatial accuracy depends on the distribution of the BTSs (or cell towers), and varies significantly. Two different groups of location data are recorded in mobile networks (Calabrese et al., 2014; Oliver et al., 2015): event-driven, and network-driven mobile phone network data.

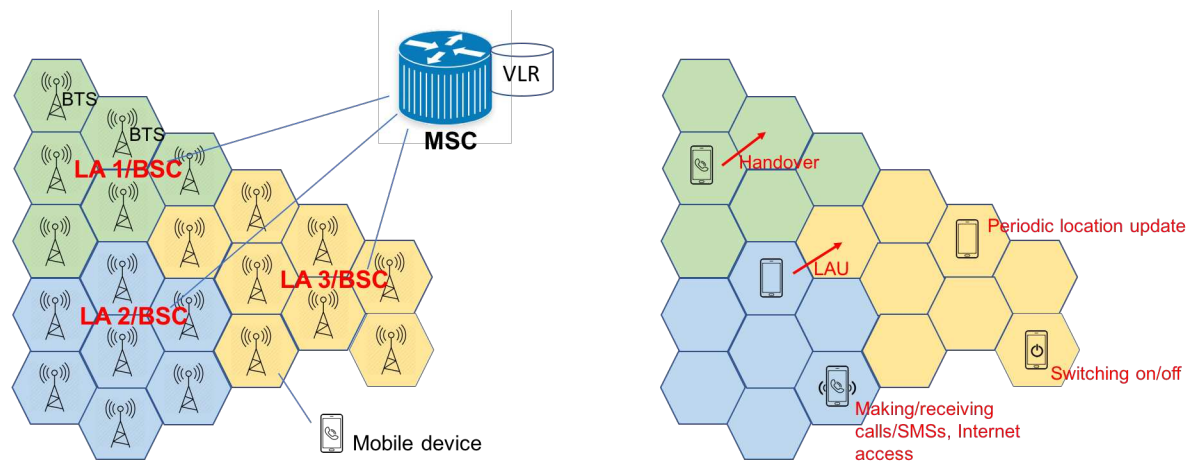


Figure 1 Mobile phone network structure (left), and network events captured in network-driven data (right).

Event-driven mobile phone network data

This group of data is recorded when the phone actively uses a service, e.g., call, short message service (SMS), or Internet access. The data are mainly for billing purposes. There are two types of data in this group:

- 1) Call Detail Records (CDRs): A CDR contains the details of a phone call or SMS passed through a phone. Its content is not necessarily standardized, and can vary between different mobile network providers. In general, a CDR typically consists of the encrypted originating and destination phone numbers, a timestamp, the duration (for calls), the communication type (call or SMS), ID of the BTS/cell the originating phone connects to, and sometimes the cell ID of the receiving phone.
- 2) Internet Protocol Detail Records (IPDRs), also known as Internet access log: An IPDR contains details of internet usage. Again, its content is determined by mobile phone operators, but typically contains the encrypted ID of the mobile phone, timestamp, number of bytes transferred, information about the website visited, and ID of the BTS/cell the phone connects to.

Network-driven mobile phone network data

To ensure the quality of communication services, mobile phone networks need to monitor locations of their subscribers. Network-driven mobile phone data (aka signaling data or sightings data) store location updates of mobile phones, mostly triggered by the following types of network events (Figure 1, right):

- 1) Switching the phone on and off: In this case, the BTS/cell where the phone was last connected is stored.
- 2) Location area update (LAU): This occurs when a standby phone moves to a cell that belongs to another LA.

- 3) Handovers: This happens when a user moves from one cell to another during a phone call or an Internet data session.
- 4) Making or receiving calls or SMS, as well as accessing Internet services: In contrast to CDRs, these events only record the location of the phone and the timestamp, with no information about the receiving phone and its location.
- 5) Periodic location update: If a phone is idle for a period of time (typically a few hours), i.e., none of the above events happen, a periodic location update request is issued to obtain the phone's location.

When comparing event-driven and network-driven mobile phone data, one can see that network-driven data are often much denser, as information is captured for all users independently of their actual use of the phone. However, network-driven data cannot be used to study human social networks, as they do not contain information regarding the other side of the communication.

Temporal and spatial granularity of mobile phone network data

The frequency of the data depends on the type of the mobile phone network data (i.e., event-driven or network-driven), and is largely user dependent (Chen et al., 2016). For example, each voice call generates one CDR. However, the same call might lead to several network-driven data records if the phone moves along multiple cells during the call. Meanwhile, a user travelling longer distances (e.g., over multiple cells) regularly leaves much more records than those who do not. Another important issue to consider is that the data are not recorded at a regular manner. A study with CDRs showed an average inter-event time of 8.2 hours for 100,000 individuals over six months (González et al., 2008). Using a network-driven dataset, Calabrese et al. (2011b) found an average inter-event time of 260 minutes, and an arithmetic average of the medians of 84 minutes. Recently, with the increasing use of mobile Internet (e.g., via the apps installed on smartphones), the temporal resolution of network-driven data has increased significantly. For example, Chin Jiaqi (2018) found that the medians of the time intervals between consecutive data points are about 90 seconds in the network-driven dataset used.

Regarding spatial accuracy, these data are significantly less accurate than GPS data (Horn et al., 2017). In most cases, the phone's location is represented by the position of the cell tower to which the phone connects (Calabrese et al., 2014), using the Cell of Origin (CoO) method. Sometimes, other methods such as timing advance (TA) and lateration/triangulation are employed to improve location accuracy, using low level measurements such as received signal strength and transmission time of multiple cell towers in available range. However, most mobile network operators do not disclose the details of how the phone's location is estimated, nor the spatial extents of each cell or LA (Widhalm et al., 2015). Previous studies showed that the spatial resolution ranges from the order of a few meters (Chen et al., 2016) to about 300 m (Calabrese et al., 2011b; Jiang et al., 2013) or 500 m (Horn et al., 2017) in urban areas where the density of cell towers is much higher, to that of several kilometers (Horn et al., 2017; Widhalm et al., 2015) in rural and less heavily populated areas.

In general, we can see that mobile phone network data (especially network-driven ones) are becoming more and more temporally frequent, and thus spatially accurate, mainly due to the increasing use of mobile Internet and the employment of more advanced positioning methods such as triangulation. However, the data quality is still not comparable to that of typical GPS data. Therefore, a fundamental question appears: Is transport mode detection from mobile phone network data even feasible? Considering that many studies have shown that it is still possible to achieve acceptable accuracy of transport mode detection using sparse GPS data (e.g., 2-minute time interval between points) (Bantis and Haworth, 2017; Bolbol et al., 2012), temporal frequency of mobile phone network data seems not to be an issue. Furthermore, the possibility of using auxiliary geographic data of the study area (e.g. stops and routes of public transport) seems to provide some good compensation of the low spatial accuracy of mobile phone network data, as shown in many existing studies with GPS data (Bantis and Haworth, 2017).

To conclude, transport mode detection using mobile phone network data (especially network-driven ones) seems to be feasible, even though it is a very challenging task.

Summary

Due to the ubiquity of mobile phones, mobile phone network data potentially allow us to analyze travel behaviors of the whole population, with full temporal coverage at a comparatively low cost. Compared to event-driven data (e.g., CDRs), network-driven data are much denser, while however not containing information about social interaction. In general, both event-driven and network-driven mobile phone network data tend to have less spatial accuracy than GPS data, and are recorded infrequently and irregularly. These characteristics make transport mode detection based on mobile phone network data very challenging. In the following sections, we analyze the state-of-the-art on this aspect, aiming to answer the questions outlined in Section 1.

3. Methodology: Systematic review

3.1 Search strategy

As mentioned before, the systematic review approach has been often employed in many scientific disciplines to minimize bias and ensure reproducibility of review studies, thus providing reliable findings from which conclusions can be drawn (Budgen and Brereton, 2006; Moher et al., 2009; Petticrew and Roberts, 2006). Following the recommendations by the PRISMA (preferred reporting items for systematic reviews and meta-analysis) statement (Moher et al., 2009) and to ensure the reproducibility of the current research study, we describe the overall peer-review process in detail in the following. Specifically, three databases, Web of Science, Scopus, and TRID (Transport Research International Documentation), were searched using keywords contained in the title, abstract, and topic. There are two categories of search terms, and at least one term from each category must be matched: 1) mobile phone network data*, mobile phone networking data*, call detail record*, CDR, cellular network, cellular phone data*, floating phone data*, phone data*, signal* data*; 2) transport*, travel, mobility. We chose rather broad terms for the second category to make sure that we do not miss any relevant papers. The search was done in January 2018, and no starting date was set. Following that, metadata (i.e., authors, title, year, abstract, and source title) of each paper was collected.

3.2 Inclusion and exclusion criteria

In order to make sure that each study included in the review is eligible, studies had to meet the following criteria: 1) written in English and published in peer-reviewed journals, conference proceedings, or books; 2) not a review paper; 3) related to the transport mode detection procedure; 4) uses data recorded by mobile telecommunication network operators (excluding cellular signaling data collected by smartphones).

The main reasons why we excluded signaling data collected by smartphones are as follows:

- 1) Mobile network data (e.g., GSM data) collected by smartphones actually are different from those collected by telecommunication operators, especially in terms of temporal frequency and spatial accuracy. For example, data collected by smartphones tend to have fixed intervals (e.g., every 10 minutes) and use the cell-of-origin (CoO) positioning method, while data collected by telecommunication operators are recorded infrequently and irregularly, and employ more positioning methods (e.g., lateration/triangulation) than CoO. In short, these two categories of data have very different temporal and spatial characteristics.
- 2) We are interested in data that cover a high percentage of the whole population. Mobile network data collected by smartphones can only cover users who install a specialized mobile application on their smartphone to collect cellular signals, which limits the scale of this type of studies. In contrast, mobile network data collected by telecommunication operators cover all

the subscribers of the operators, which normally consist of a very high percentage of the whole population.

After this eligibility check, 13 papers remained in the list. We then further checked the references cited by these 13 papers, and those that cited these papers after their publication using Google Scholar (<https://scholar.google.com/>). We applied the same eligibility check to these references and citations. 6 papers were further identified. For the previously published review papers, we examined their references, and added those missing but relevant papers to the paper list. This step was finished in May 2018. In total 22 papers matching all the criteria were included in this review. Figure 2 shows the search and screening process.

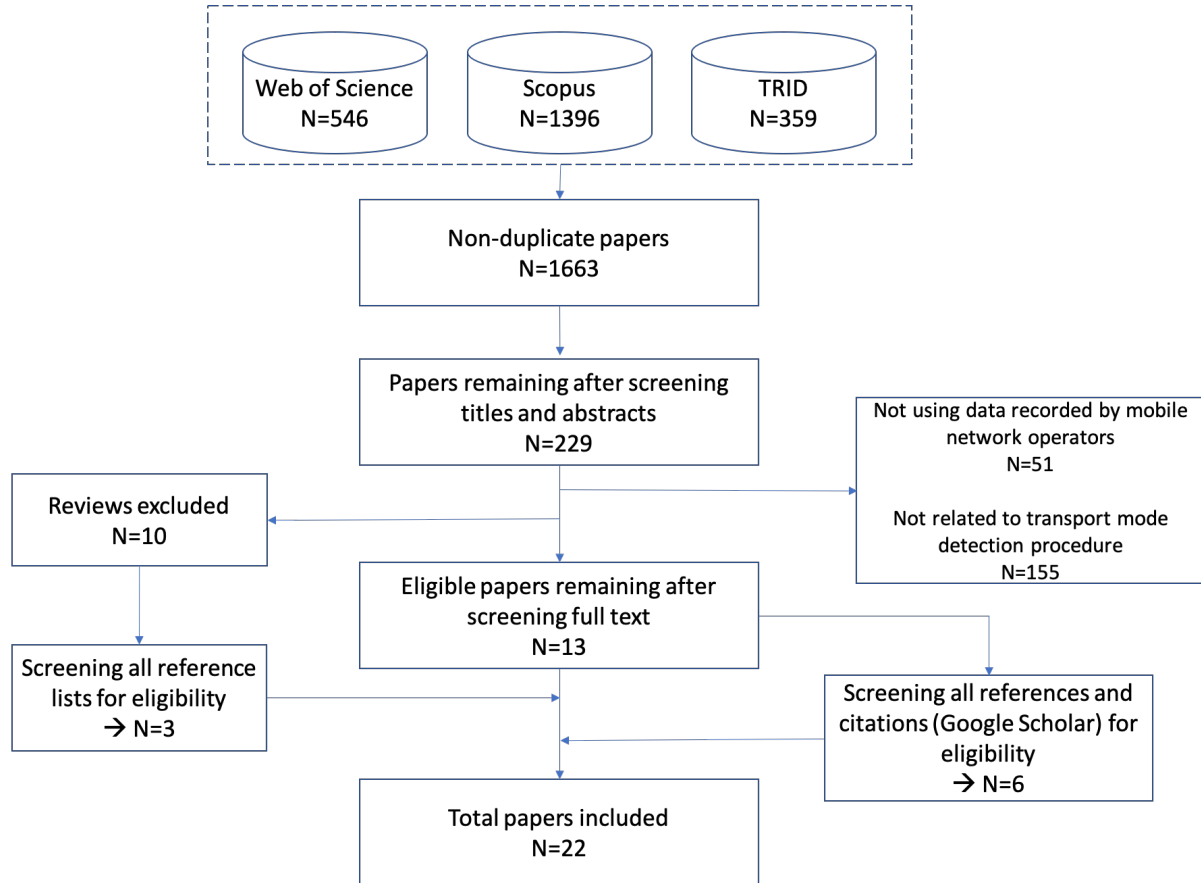


Figure 2. Flowchart of the systematic review process

3.3 Data extracted

The information extracted from each eligible paper included:

- 1) Data: data types (e.g., event-driven and network-driven) and data characteristics (e.g., location determination methods, spatial and temporal accuracy)
- 2) Pre-processing: outlier/noise removal, trip identification, and features and thresholds used for trip identification
- 3) Mode detection: transport modes detected, additional data used (e.g., geodata of railway network), mode detection algorithms (e.g., rule-based methods, unsupervised and supervised methods), and attributes/features employed
- 4) Validation/evaluation: ground truth data used, evaluation methods, and performance of mode detection methods
- 5) Other specific aspects of the study

4. Results

This section summarizes the information extracted from the 22 eligible papers remaining after the systematic review process of Section 3. We particularly focus on the following aspects: data used and their characteristics, data pre-processing (data cleaning and segmentation), and transport model detection (algorithms and validation methods).

4.1 Data and their characteristics

As can be seen from Section 2, different types of mobile phone network data are recorded by telecommunication network operators. A summary of the data used in the eligible papers is presented in Table 1.

Table 1 Summary of mobile phone network data used in the eligible studies

Type of data	Publications
Event-driven data (n=7)	Wang et al. (2010), Doyle et al. (2011), Wu et al. (2013), Qu et al. (2015), García et al. (2016), Kalatian and Shafahi (2016), Phithakkitnukoon et al. (2017)
Network-driven data (n=15)	Schlaich et al. (2010), Calabrese et al. (2011), Xu et al. (2011), Smoreda et al. (2013), Horn and Kern (2015), Larijani et al. (2015), Holleczeck et al. (2015), Asgari (2016), Poonawala et al. (2016), Yamada et al. (2016), Danafar et al. (2017), Li et al. (2017), Hui et al. (2017), Hui (2017), Horn et al. (2017)

As can be seen from Table 1, 15 out of the 22 studies (about 70%) used network-driven data, while the other 7 studies used event-driven data, particularly CDRs. In terms of location estimation, cell of origin method was often mentioned.

Surprisingly, only three studies (Kalatian and Shafahi, 2016; Larijani et al., 2015; Wang et al., 2010) mentioned the characteristics of the data used, and reported that the location accuracy was around 200-400m in urban areas. All the other studies did not mention the data characteristics, e.g., location accuracy and temporal frequency, as well as what kinds of network events (e.g., handovers) are included when network-driven data are used. A possible reason might be that most mobile network operators (i.e., data providers) do not disclose the details of their location estimation method, as well as the spatial extent of each cell. The missing information regarding location accuracy might also be due to the fact that there was no reference dataset (e.g., corresponding GPS data) available for comparison.

4.2 Data cleaning

Mobile phone network data are characterized by spatio-temporal uncertainties (Wang and Chen, 2018), which require some data cleaning procedures before performing the follow-up data analysis. The data cleaning process typically consists of removing short or incomplete trajectories that only contain few data points, interpolation of missing data points (during long time gaps), and filtering of outliers and noise considering their temporal and spatial characteristics.

In terms of temporal uncertainties, depending on the device usage pattern and mobility, the temporal distribution of location records can be irregular and heterogeneous. The average time interval between two consecutive records could be too long to capture a user's movements that took place during the time gap. A common way to address this temporal issue is to filter out users with a low number of recorded events (Calabrese et al., 2013; Iovan et al., 2013; Zhao et al., 2016). There are also efforts of using interpolation methods to recover missing data points between two consecutive records with a long time gap (Järv et al., 2017; Louail et al., 2014).

From the spatial perspective, oscillation (aka ping-pong effect) is the most prominent phenomenon that brings noise and outliers to mobile phone network data. It occurs when the mobile device

switches its connection between multiple BTSs (typically within a short time interval) even though the device itself is not moving (Calabrese et al., 2011a; Iovan et al., 2013; Miao et al., 2016). This might be due to fluctuations in the received signal strength as well as load balancing policies of the network. The oscillation phenomenon generates a considerable number of records that do not reflect actual device movements (e.g., 30% in Lee and Hou (2006)). When triangulation/TA methods (instead of cell of origin) are used for location estimation, the fluctuations in the signal strength also have a strong impact on the data quality, leading to different estimates for the same location (Calabrese et al., 2011a; Wang and Chen, 2018).

Three types of methods have been proposed to address the oscillation issue in mobile phone network data: pattern-based, speed-based, and hybrid methods. Pattern-based methods extract location sequences that exhibit specific switching patterns (e.g., $L_0-L_1-L_0-L_1-L_0$, where L_0 and L_1 are IDs of cell towers) using heuristic rules, and label them as oscillation sequences (Bayir et al., 2010; Lee and Hou, 2006; Shad et al., 2012). These oscillation sequences are then removed. These methods often require the information about which cell towers the phone is connected to, particularly the ID of the cell tower. Meanwhile, these methods heavily rely on heuristic rules, which often have difficulties to detect complex oscillation cases and might lead to mistakenly removing actual travels. Based on the observation that oscillation usually leads to an incredibly high switching speed (e.g., 500 km/h), speed-based methods identify an oscillation sequence if the switching speeds between location records within a time window exceed a given threshold (Horn et al., 2014; Iovan et al., 2013). Some other information might be also used, such as heading changes. Hybrid methods combine pattern-based and speed-based methods to identify complex oscillation patterns and reduce the risk of mistakenly removing real trips (Wang and Chen, 2018; Wu et al., 2014).

Table 2 summarizes the data cleaning methods reported in the eligible papers. Surprisingly, among the 22 papers, only 5 reported how they cleaned their data before other data analysis. These studies mainly focused on filtering of outliers and noise from the data, and none of them dealt with the issue of interpolating/recovering missing data points. Schlaich et al. (2010) and Li et al. (2017) applied pattern-based methods to address the oscillation issue, using several simple heuristic rules. These methods work with CoO location data, and can detect simple oscillation cases. However, they often fail to detect complex oscillation cases and sometimes mistakenly remove real trips, due to the difficulty of developing comprehensive heuristic rules.

Considering temporal and spatial constraints, Kalatian and Shafahi (2016) aggregated a set of nearby cells that present fluctuations in the raw data as a cell cluster, and used a weighted center of the cell to replace these cells. Horn and Kern (2015) and Horn et al. (2017) employed a recursive look-ahead filter previously developed by Horn et al. (2014). First, the speed between two consecutive records O_{i-1} and O_i is calculated and if this speed exceeds a threshold $V_{supersonic}$ (260 km/h in their case, which is twice the maximum speed limit of an Austrian highway), O_i is labelled as a potential outlier. The look-ahead portion of the filter then calculates the distance between O_i and O_{i+1} , $D_{i, i+1}$, and the distance between O_{i-1} and O_{i+1} , $D_{i-1, i+1}$. If $D_{i, i+1}$ is larger than $D_{i-1, i+1}$, O_i is considered as the outlier and removed. Otherwise, O_{i-1} is removed.

Table 2 Summary of eligible papers reporting data cleaning methods (n=5)

Publication	Data type	Location estimation	Data cleaning method
Schlaich et al. (2010)	Network-driven	Cell of Origin (CoO)	Pattern-based
Horn and Kern (2015)	Network-driven	Triangulation	Speed-based
Kalatian and Shafahi (2016)	Event-driven (CDR)	CoO	Hybrid
Horn et al. (2017)	Network-driven	Triangulation	Speed-based
Li et al. (2017)	Network-driven	CoO	Pattern-based

It is important to note that none of the studies above has evaluated their data cleaning methods, probably due to a lack of reference dataset (e.g. corresponding GPS trajectories). As the recorded trajectory represents a movement in space, which is often restricted by the underlying geographic context (e.g., road network), map matching methods might be also employed to further improve the performance of data cleaning. Map matching is a very popular data cleaning method employed for GPS data processing (Newson and Krumm, 2009).

4.3 Trip identification

After cleaning the raw data, the next step of transport mode detection is to identify trips (more precisely, trip stages), each of which normally contains a single transport mode. This step is often called trip identification (TI) or segment identification. To achieve this, key places might be identified, and differentiated on whether these places are meaningful stops or the user is merely passing through them. The former, also known as stay locations, potentially represent start and end locations of a trip. A summary of the TI methods and the features employed in the eligible papers is shown in Table 3.

Table 3 Summary of eligible papers reporting information about trip identification methods and the features employed (n=17)

Publication	Location estimation	TI Method	Features and thresholds used
Schlaich et al. (2010)	Cell of Origin (CoO)	Rule-based method	Stay duration at a location area LA, duration threshold (= 60 min)
Yamada et al. (2016)	CoO	Rule-based method	Present frequency at a cell within a fixed time window, stay duration threshold (= 30 min), frequency ratio (= 0.5)
Kalatian and Shafahi (2016)	CoO	Rule-based method	Stay duration at a cell
Wang et al. (2010)	triangulation	Spatio-temporal clustering	Spatial and temporal distance between records, spatial threshold (= 1 km), time threshold
Poonawala et al. (2016)	?	Spatio-temporal clustering	Spatial and temporal distance between records, spatial threshold, time threshold
Qu et al. (2015)	?	Frequency-based method (for commuting trips)	Present frequency at a census tract at specific time periods (8:00pm-7:00am for home, 9:00am-5:00pm on weekdays for work location)
Phithakkitnukoon et al. (2017)	CoO	Frequency-based method (for commuting trips)	Present frequency at a cell at specific time periods (10:00pm-7:00am for home, 9:00am-5:00pm on weekdays for work location)
Doyle et al. (2011)	CoO	Georeferencing (for intercity trips)	Geographic boundary of target cities
Wu et al. (2013)	CoO	Georeferencing (for intercity trips)	Geographic boundary of target cities
Horn and Kern (2015)	triangulation	Georeferencing (for trips between train stations)	Geographic location of target train stations, buffer size (= 1 km)
Larijani et al. (2015)	CoO	Georeferencing (for trips between Voronoi cells)	Geographic location of cell towers
Holleczeck et al. (2015)	CoO	Georeferencing (for trips between MRT/subway stations)	List of indoor cell towers (which serve exclusively MRT passengers)
García et al. (2016)	CoO	Georeferencing (for intercity)	Geographic boundary of target

		trips)	cities
Li et al. (2017)	CoO	Georeferencing (for trips between MRT stations)	Geographic location of target MRT stations
Hui et al. (2017)	CoO	Georeferencing (for intercity trips)	Geographic boundary of target cities
Hui (2017)	CoO	Georeferencing (for intercity trips)	Geographic boundary of target cities
Horn et al. (2017)	triangulation	Georeferencing (for trips between traffic analysis zones TAZs)	Geographic boundary of TAZs, stay duration at a TAZ, duration threshold (= 60 min)

As can be seen from Table 3, the methods used for trip identification in the eligible papers can be classified into the following types (with number of occurrences in brackets): rule-based methods (n=3), frequency-based methods (n=2), georeferencing (n=10), and spatio-temporal clustering (n=2). Georeferencing has been employed most often, followed by rule-based methods, and lastly frequency-based methods and spatio-temporal clustering.

- **Georeferencing** (41% of the 22 eligible studies) makes use of the geographic data of the study area, and simply checks whether a raw trajectory intersects the geographic boundary (or a buffer) of a list of pre-defined spatial areas, such as cities, train stations, MRT/subway stations, and traffic analysis zones (TAZs). In the eligible papers, this approach has been applied for identification of intercity trips (Doyle et al., 2011; García et al., 2016; Hui et al., 2017; Hui, 2017; Wu et al., 2013), trips between MRT stations (Holleczek et al., 2015; Li et al., 2017), trips between train stations (Horn and Kern, 2015), and trips between TAZs (Horn et al., 2017). To ensure that an intersected spatial area is a stay location (instead of a pass-by location), Horn et al. (2017) further checked the duration of stay within the area against a given threshold (60 minutes). In general, georeferencing works well for cases when the trips only start and end from a list of pre-defined spatial areas.
- **Rule-based methods** (14%) are built on the assumption that if a user remains a considerably longer time in a cell or LA, the user potentially starts or ends a trip in the respective cell or LA. A duration threshold (or similar) is also used. These rules are human-crafted or curated, mostly making use of common sense or prior knowledge. For example, Schlaich et al. (2010) suggested a 60-min rule in which, if the time period between a login and its corresponding logout of a visited LA is 60 minutes or more, then it is considered to be a stay location, and marks the end of a previous trip and the start of a new trip. Kalatian and Shafahi (2016) set the duration threshold according to the largest diameter of a cell. To deal with noise and outliers, Yamada et al. (2016) identified the BTS/cell which the user is connected to most frequently in a fixed time window, and checked whether the user's stay period in this cell is longer than a threshold (30 minutes in their paper). If yes, a stay location is identified.
- **Frequency-based methods** (9%) assume that the most visited cells are connected to important places in a person's life. Very often, visiting frequencies within different time windows (e.g., nighttime, daytime) are computed to detect a person's home or work location. Two of the eligible papers (Phithakkitnukoon et al., 2017; Qu et al., 2015) used these methods to identify commuting trips. Frequency-based methods work well for data with cell-level locational quality, and long observation periods (at least multiple days).
- **Spatio-temporal clustering** (9%) tries to identify stay locations using both spatial and temporal constraints, addressing the noisy and raw nature of mobile phone network data. Typically, this approach first groups consecutive location points that are spatially close by measuring the distances between points and comparing them to a distance threshold. This step can be done using existing density-based clustering algorithms such as DBSCAN (Ester et al., 1996). In a second step, each cluster obtained is checked to see whether the time difference between the first and the last observation in the cluster exceeds a time threshold. If yes, the medoid, centroid or other weighted location of the cluster is computed and the

original points in the cluster are deleted. These medoids/centroids are then set as the start or end points of each trip. Two of the eligible papers (Poonawala et al., 2016; Wang et al., 2010) used spatio-temporal clustering to identify trips. In general, this approach can detect stay locations despite noise and outliers presented in the dataset. However, the spatial and time thresholds should be carefully set.

Table 4 compares the strengths and weaknesses of these methods, mainly regarding the following aspects: data types supported (cell-of-origin (CoO) location data, or location data estimated by triangulation/TA); whether data over a long period (at least multiple days) are needed; use of GIS data (e.g., geographic boundary of train stations); and resilience to noise and outliers.

Table 4 Strengths and weaknessess of the trip identification methods employed

Method	Data type	Long period data needed?	GIS data needed?	Resilience to outliers	Notes
Georeferencing	CoO and triangulated	No	Yes	Fair	Works well when the trips between pre-defined areas
Rule-based	CoO	No	No	Poor	Requires heuristic rules
Frequency-based	CoO	Yes	No	Fair	Often for identifying commuting trips
ST clustering	CoO and triangulated	No	No	Good	Often based on distance and time thresholds

As can be seen from Table 4, georeferencing methods work well when the trips to be identified start and end in a list of pre-defined areas, e.g., trips between train/metro stations, and intercity trips. Frequency-based methods are often employed to identify frequent trips, e.g., home-work commuting trips. Compared to the other three types of methods, spatio-temporal clustering methods can much better deal with noise and outliers in the data, by considering both spatial and temporal constraints. Current studies employing spatio-temporal clustering for trip identification require to manually set values for the distance and time thresholds. However, they can also be automatically learned from the data by applying a model-based clustering method (Chen et al., 2014).

4.4 Transport mode detection

Once trips are extracted, they are now ready to be analyzed to identify the trip modes. In order to do so, information on each trip is extracted as trip features/attributes, which are then used in a mode detection algorithm to assign the transport mode(s) to the trip. In this section, we summarize and analyze the mode detection methods used in the eligible papers, particularly focusing on the following aspects: transport modes to be detected, additional data used (e.g., geographic data or timetable of public transport), features/attributes employed in the mode detection algorithms, detection algorithms, spatial analysis techniques for measuring spatial proximity (if any), and datasets used for evaluation or validation (if any). Table 5 summarizes the results.

4.4.1 Modes detected, additional data used, and attributes extracted

Transport modes detected. The number of modes detected in the eligible papers ranged from 1 to 6, with 16 papers trying to differentiate between 2 or 3 modes. Train and car were the two most popular modes being detected. Many of the eligible papers focused on detecting modes of intercity trips (Doyle et al., 2011; García et al., 2016; Hui et al., 2017; Hui, 2017; Schlaich et al., 2010; Smoreda et al., 2013; Wu et al., 2013), which seem to be easier to infer considering the low spatial and temporal resolution of mobile phone network data. Qu et al. (2015) and Phithakkitnukoon et al. (2017) investigated modes of commuting trips, exploiting users' trajectory data over a period of time.

Some of the eligible papers (Holleczek et al., 2015; Horn and Kern, 2015; Li et al., 2017; Wu et al., 2013) focused on easy-to-detect modes (e.g., metro, train or plane), while ignoring those difficult to infer (e.g., bike, bus, tram). There were also many studies aggregating different modes into more general groups, such as public transport versus private transport (Horn et al., 2017; Phithakkitnukoon et al., 2017; Qu et al., 2015; Wang et al., 2010), air versus ground (Hui et al., 2017; Hui, 2017), moving versus stationary (Calabrese et al., 2011a), and rail versus road (Asgari, 2016; Doyle et al., 2011). The use of more general mode groups is mainly due to the low spatial and temporal resolution of mobile phone network data. Danafer et al. (2017) is the only study trying to differentiate 6 different modes: car, bus, tram, train, cycling and walking. However, no accuracy of their proposed detection algorithm has been reported.

Additional data use. Regarding the use of additional data (beyond the trip data whose modes are to be inferred), 16 out of the 22 eligible papers made use of the geographic data of the underlying transport networks (e.g., railway, metro network), road networks, or airports. This matches our expectation, as humans and particularly vehicles are chiefly bound to moving in a geographic network. Yamada et al. (2016) and Horn et al. (2017) further considered train timetables/schedules to improve the performance of their mode detection algorithms.

Attributes/features extracted. In terms of the attributes/features used in the mode detection algorithms, “proximity to network” was the most popular one, especially for the papers making use of geographic data. For more details on how this feature was computed and used, please refer to Section 4.4.3. In addition to “proximity to network”, the following trip level attributes were often employed: trip duration (i.e., time gap between the first and last data points), trip distance (i.e., distance covered by all data points) and trip speed (mean and standard deviation).

4.4.2 Transport mode detection algorithms

As can be seen from Table 5, the mode detection algorithms proposed in the eligible papers can be classified into three groups: rule-based heuristics (RBH), clustering (unsupervised machine learning), and statistical analysis. Among them, RBH has been employed in most of the eligible papers.

Rule-based heuristics (RBH). Methods using RBH often involve human-crafted or curated rule sets, mostly making use of prior knowledge (common sense or expert knowledge). These rules either compare features/attributes extracted to pre-defined thresholds (e.g., if the average speed of a trip is bigger than 300 km/h, it is likely a plane trip), or give preferences to the one with the best “match” (e.g., if a trip is closer to the railway network than that to the road network, it is more likely a train trip). For example, Hui et al. (2017) employed two simple rules for inferring modes of trips between two cities in Canada: 1) If trip duration is between 0.5-1.5h, assign as an air trip; 2) if trip duration is between 2-6 h, assign as a ground trip. A relatively more comprehensive RBH example can be found in Qu et al. (2015). They detected transport modes from CDR data using an RBH method that combines trip speed, trip distance, proximity to public transport network, and a logit model. The rules employed were: 1) If travel speed of a trip is higher than 15 km/h, and no subway or bus stations are within 500 meters of its origin or destination, the trip is assigned a car mode; 2) if trip speed is lower than 8 km/h, and trip distance is 0 to 3 km, the trip is assigned a walking mode; 3) all other cases are determined by a logit model, which predicts an individual’s choice based on utility or attractiveness of a particular mode.

18 out of the 22 eligible papers employed RBH to detect transport modes of trips. Hui et al. (2017) and Calabrese et al. (2011a) simply used attributes such as trip duration or trip speed. All the other RBH methods considered spatial proximity to public transport network or road network in their rules. Among them, six papers only employed spatial proximity for mode detection (Doyle et al., 2011; Holleczech et al., 2015; Horn and Kern, 2015; Phithakkitnukoon et al., 2017; Poonawala et al., 2016;

Wu et al., 2013), which were mostly for intercity trips or trips between metro/MRT/train stations. The other 10 papers additionally considered other trip-related attributes, such as trip speed (Danafar et al., 2017; Larijani et al., 2015; Smoreda et al., 2013; Yamada et al., 2016), trip distance (Asgari, 2016; Qu et al., 2015), trip duration (García et al., 2016; Li et al., 2017; Schlaich et al., 2010), and temporal overlap with timetables of public transport (Horn et al., 2017; Yamada et al., 2016).

Different from the other studies that employed RBH, Calabrese et al. (2011a) learned the trip speed threshold from labeled data, instead of setting the threshold based on common sense or prior knowledge.

Clustering (unsupervised machine learning). Clustering methods, particularly k-means and hierarchical agglomerative clustering, were employed in 3 of the eligible papers for transport mode detection. These methods group “unlabeled” data (i.e., trips without their modes assigned) into different clusters, based on key features/attributes of the data (e.g., duration and travel speed of each trip). The mode of the trips in each cluster is then assigned by analyzing the characteristics of its centroid or mediod, making use of common sense, prior knowledge, or other data sources (e.g., travel time obtained from Google Maps).

Wang et al. (2010) is the first study in this class. They used the k-means method to cluster trips based on their duration. The method partitioned all the trips into two separate clusters corresponding to the modes of interest, namely private (driving) and public transport. The cluster with a lower average trip duration (and similar to driving time reported by Google Maps) was assigned driving, and vice versa. Kalatian and Shafahi (2016) presented another k-means classification, which instead used trip speed. They differentiated between walking, private car and bus modes. Instead of using k-means, Hui (2017) employed a hierarchical agglomerative clustering method to cluster trips based on their duration. The optimal number of clusters was determined by applying the R package NbClust (Charrad et al., 2014) and considering the modes of interest. The clusters were then assigned corresponding modes by checking the average trip duration of each cluster.

Statistical analysis. Xu et al. (2011) developed a probabilistic method for inferring the transport modes driving, biking and walking. Their aim was to infer the mostly likely mode sequence, given the current trip and some historical data of relevant trips. The proposed method consisted of a Hidden Markov Model (HMM) with two sub-models for different traffic conditions. The authors used speed distribution law to detect transport modes under the normal state, in which the parameters (speeds of biking and walking) were learned from the (labeled) training set. For the congested condition, Cumulative Prospect Theory was used to improve the performance of mode detection.

Comparing these three groups of mode detection algorithms, we can see that RBH methods rely on the existence of some prior understanding of the transportation modes to be detected, at least in the form of common sense rules. Together with data about the geographic environment (i.e., geodata), they work well for detecting trips along railways (e.g., train, commuter train, metro). They also work well for differentiating trips whose modes present significantly different characteristics, e.g., air versus ground. The difficulty here is to develop proper rules as well as finding appropriate threshold values, particularly with modes that exhibit high intra-class variation, leading to overlaps between classes. Clustering methods try to separate all trips into different groups, and therefore require the availability of other trips (in addition to the current one whose mode is to be detected). Meanwhile, assigning modes to the resulting clusters is often done manually (i.e., by human interpretation), and might not be an easy task if differences between clusters are very small, or do not match prior knowledge.

4.4.3 Spatial analysis techniques for measuring spatial proximity

As mentioned before, spatial proximity to public transport and the road network as well as other spatial objects was the most popular feature employed for transport mode detection in the eligible papers. The assumption is that if a certain transport means was used, the recorded trajectory is spatially more “proximate” to its underlying network than to the other network, e.g., the average distance is shorter. In the following, we analyze how geographic data were used to measure spatial proximity in the eligible papers. Three groups of methods can be differentiated: methods based on Euclidean distance to network (n=6), methods based on connection to specific cell towers (n=7), and map matching (n=3).

The first group of methods computes a trip’s (Euclidean) distance to the underlying networks or stations (Horn et al., 2017; Horn and Kern, 2015; Phithakkitnukoon et al., 2017; Qu et al., 2015; Smoreda et al., 2013; Yamada et al., 2016). For example, Horn et al. (2017) assigned a trip with the mode whose underlying network has a shorter average distance to the trip. Qu et al. (2015) checked whether the origin or destination of a trip is within a 500 m buffer of a subway or bus station. They simply used the results to differentiate between car mode and public transport.

The second group of methods checks whether a trip (i.e., a trajectory) has more connections to specific cell towers, e.g., along the railway or highway network, respectively. This group is often used with CoO based location data, and often requires the availability of cell tower locations. A common approach in this group consists of two steps: 1) Identify a set of cell towers for each mode of interest (e.g., cell towers along a railway network for train mode), by checking whether the underlying network crosses the coverage of a cell tower; 2) check which set has more overlap with the trip trajectory, and assign the transport mode accordingly. Very often, the coverage of a cell tower used in the first step was approximated by a Voronoi polygon (Doyle et al., 2011; García et al., 2016; Wu et al., 2013). Larijani et al. (2015) and Holleczeck et al. (2015) made use of their specific network setting to identify a set of cell towers for Step 1. In their studies using data from Paris and Singapore, dedicated cell towers that exclusively serve metro or MRT users had been installed by mobile network operators. Therefore, they checked whether a trip has records linked to these dedicated cell towers to identify metro or MRT modes.

The third group of methods uses map matching techniques to map a trip to the underlying network. Two different map matching techniques have been applied: Hidden Markov Model (HMM) (Asgari, 2016; Poonawala et al., 2016), and Bayesian inference (Danafar et al., 2017). For example, Asgari (2016) applied HMM-based map matching to find a likely road path and a likely rail path for the original trip trajectory, and checked which of the two paths has a better match with the original trajectory.

4.4.4 Evaluation and validation of mode detection methods

Surprisingly, 10 out of the 22 eligible papers did not report how the proposed mode detection methods are evaluated or validated. Several of the papers employed mode share statistics data from official sources or self-reports for validation (García et al., 2016; Phithakkitnukoon et al., 2017; Qu et al., 2015; Wang et al., 2010). They mainly compared the difference between the percentage of trips in each mode obtained from the mobile phone network data, with the percentage from mode share statistics. Hui et al. (2017) and Hui (2017) validated their methods by checking whether the extracted air-travel trips matched with flight routes. There were also studies validating their methods with aggregated data obtained from manual counting (Holleczek et al., 2015) or data sources such as farecard data (Poonawala et al., 2016).

Compared to the above studies, which used aggregated data for evaluation or validation, there were four papers evaluating their methods with individual ground truth data, either simulated (Yamada et

al., 2016), or labeled by users (Asgari, 2016; Li et al., 2017; Xu et al., 2011). For example, Yamada et al. (2016) used a dataset simulated using the Scenargie software (<https://www.spacetime-eng.com>). For train mode, they found a recall over 0.9 in trips with more than 15 location records, and a precision of 0.9 for all cases. For car mode, the average precision and recall were 0.7 and 0.75, respectively. Xu et al. (2011) evaluated their method with 500 mode sequences labeled by individual users, and used 80 % of the data for training and the other 20 % for testing. They found F1 values of 0.89-0.93, depending on the parameters of the proposed method. Asgari (2016) evaluated their method with data labeled by users, and found a F1 value of 0.83 when differentiating road from rail mode. Li et al. (2017) used a labeled dataset consisting of 7 days data of 10 users to evaluate their method, which detected MRT trips. They defined the precision of a detected trip as the ratio of the number of correctly detected stations and the total number of stations in the detected trip, and recall of a real trip as the ratio of the number of correctly detected stations and the total number of stations in the real trip. The reported average precision, recall and F1 values were around 0.64-0.91, 0.74-0.81, and 0.72-0.83, respectively, depending on the parameters.

In summary, due to the small number of studies that reported how well their proposed methods were able to detect modes of individual trips, as well as the lack of a “benchmark” labeled dataset and standardized evaluation procedure, it is difficult to draw conclusions regarding which mode detection methods perform best.

Table 5 Summary of transport mode detection methods and extracted attributes in the eligible papers

Publication	Transport modes	Additional data used	Attributes	Spatial proximity analysis technique	Mode detection algorithm	Validation dataset
Rule-based Heuristics RBH (based on prior knowledge) (n=18)						
Hui et al. (2017)	Intercity modes: plane, ground	-	Trip duration	n/a	RBH (rule-based heuristics)	Flight data
Calabrese et al. (2011a)	Moving (fast mode), stationary/walking	-	Trip speed	n/a	RBH (rule-based heuristics, thresholds learnt from labeled data)	n/a
Smoreda et al. (2013)	Intercity modes: plane, train, car	Geodata of railway and road network, and airports	Proximity to network, trip speed	Euclidean distance to network	RBH (rule-based heuristics)	n/a
Horn and Kern (2015)	Train	Geodata of train station (location)	Proximity to train stations	Euclidean distance to network	RBH (rule-based heuristics)	n/a
Qu et al. (2015)	Commuting modes: private car, public transport, walking	Geodata of transport and road network	Trip speed, proximity to network, trip distance	Euclidean distance to network	RBH (rule-based heuristics)	Mode share statistics
Yamada et al. (2016)	Train, car	Geodata of railway and road network, train timetable	Proximity to network and schedule, trip speed	Euclidean distance to network	RBH (rule-based heuristics)	Simulated data
Horn et al. (2017)	train, private transport (cars/bikes)	Geodata of railway and road network, train timetable	Proximity to network and train schedule	Euclidean distance to network	RBH (rule-based heuristics)	n/a
Phithakkitnukoon et al. (2017)	Commuting modes: public transport, private car	Routes obtained from Google Maps	Proximity to routes Google Maps	Euclidean distance to network	RBH (rule-based heuristics)	Mode share statistics
Schlaich et al. (2010)	Intercity modes: cars, trucks (slow cars), other (w/ stops)	Geodata of road network	Proximity to network, trip duration	Connection to specific cell towers	RBH (rule-based heuristics)	n/a
Doyle et al. (2011)	Intercity modes: train (railway), car (road)	Geodata of railway and road network	Proximity to network	Connection to specific cell towers	RBH (rule-based heuristics)	n/a
Wu et al. (2013)	Intercity modes: car (highway)	Geodata of highway network	Proximity to highway	Connection to specific cell towers	RBH (rule-based heuristics)	n/a

Holleczeck et al. (2015)	Mass Rapid Transit MRT (metro)	Geodata of MRT network, a list of dedicated MRT cell towers	Proximity to network	Connection to specific cell towers	RBH (rule-based heuristics)	Manual counting
Larijani et al. (2015)	Metro, commuter train, car	Geodata of transport network, a list of dedicated metro cell towers (exclusively serving metro)	Proximity to network, trip speed	Connection to specific cell towers	RBH (rule-based heuristics)	n/a
García et al. (2016)	Intercity modes: train, plane, road (e.g., car)	Geodata of transport network and road network	Proximity to network, trip duration	Connection to specific cell towers	RBH (rule-based heuristics)	Mode share statistics
Li et al. (2017)	Mass Rapid Transit MRT	Geodata of MRT network, travel time between MRT stations	Proximity to network, trip duration	Connection to specific cell towers	RBH (rule-based heuristics)	Labelled data
Asgari (2016)	Rail(train/metro/tram), road	Geodata of transport network and road network	Trip distance, proximity to network	Map matching	RBH (rule-based heuristics)	Labelled data
Poonawala et al. (2016)	MRT (train)	Geodata of railway network	Proximity to network	Map matching	RBH (rule-based heuristics)	OD matrix data
Danafar et al. (2017)	Car, bus, tram, train, cycling, walking	Geodata of transport network and road network	Trip speed, proximity to shortest routes	Map matching	RBH (rule-based heuristics)	n/a
Clustering (unsupervised machine learning) (n=3)						
Wang et al. (2010)	Public transport, private car	Travel time obtained from Google Maps	Trip duration	n/a	k-means	Mode share statistics
Kalatian and Shafahi (2016)	Walking, private car, bus	n/a	Trip speed	n/a	k-means	Mode share statistics
Hui (2017)	Intercity modes: plane, ground w/o stop, ground w/ stop	n/a	Trip duration	n/a	Hierarchical agglomerative clustering	Flight data
Statistical analysis (n=1)						
Xu et al. (2011)	Driving, biking, walking	n/a	Mean and SD of trip speed	n/a	HMM (supervised)	Labelled data

5. Discussion

5.1 Summary

Summarizing the above analysis of the eligible studies of this systematic review, we can see that:

- 1) Network-driven data have been used more often for transport mode detection, with about 70% of the studies employing them. However, there is not clear trend regarding which type was preferred overtime. Moreover, the reviewed studies seldom reported the characteristics of the data, such as the location accuracy and temporal frequency, probably due to a lack of reference dataset (e.g., corresponding GPS data) available for comparison.
- 2) Only a few studies reported how they cleaned their data, especially regarding the oscillation issue; Three types of data cleaning methods have been employed to remove oscillation: pattern-based, speed-based and hybrid.
- 3) In terms of trip identification, existing studies employed four types of approaches: rule-based, frequency-based, georeferencing, and spatio-temporal clustering. Among them, georeferencing, which makes use of the geographic data of the study area, has been employed in most of the studies.
- 4) Most of the studies focused on easy-to-detect modes (e.g., metro, train or plane), and ignore those difficult to infer (e.g., bike, bus tram). Due to the low spatial and temporal resolution of mobile phone network data, many studies aggregated fine-grained modes into more general groups, e.g., public versus private transport, air versus ground. Three groups of transport mode detection algorithms were employed in the studies: rule-based heuristics (RBH), (unsupervised) clustering, and statistical analysis (using a hidden Markov model). Among them, RBH was the most popular one, used in 80% of the studies.
- 5) “Spatial proximity to network” was the most popular feature used for transport mode detection, requiring the use of auxiliary geodata. To measure spatial proximity, methods based on Euclidean distance to network, methods based on connection to specific cell towers, and map matching methods were developed.
- 6) Due to the lack of ground truth data, existing studies either did not validate their results, or simply validated their proposed methods with aggregated data such as mode share statistics.

5.2 Comparison with GPS-based transport mode detection

Several main differences can be seen by comparing the state-of-the-art of transport mode detection based on GPS data and other smartphone sensor data (Prelipean et al., 2017; Wu et al., 2016), and mobile phone network data. Firstly, the transport modes detected from GPS data (and other sensor data such as accelerometer data) tend to be more fine-grained, while - as mentioned in Section 5.2 - studies using mobile phone network data mainly focused on coarse-grained modes (e.g., rail versus road, public versus private transport) or easy-to-detect modes (e.g., metro, and train, which follow dedicated network). Secondly, supervised machine learning ML methods (e.g., Random Forest, Decision Trees, Support Vector Machine, and (Deep) Neural Networks) (Bantis and Haworth, 2017; Dabiri and Heaslip, 2018; Prelipean et al., 2016) and fuzzy logic-based methods (Schuessler and Axhausen, 2009), which are very popular in GPS-based transport mode detection, have not yet been considered for mode detection with mobile phone network data. This absence of ML methods is mainly due to the lack of labelled data. Thirdly, more features beyond “proximity to network”, such as accelerations, direction changes (turning angle), and percentile (instead of pure mean) of the extracted attributes, have been employed for GPS-based mode detection. Lastly, most of the studies on transport mode detection with mobile phone network data did not evaluate how well their proposed methods were able to detect modes of individual trips, which is actually a “default action” in research on mode detection with GPS or other smartphone sensor data.

5.3 Research gaps

From the above state-of-the-art analysis, we can see that while substantial amount of efforts have been spent on transport mode detection using mobile phone network data, there is no commonly accepted approach. Several major research gaps can be identified.

- **Data cleaning:** Currently, only few simple methods have been proposed to remove noise or outliers prior to the actual mode detection process. Moreover, evaluation of these methods is still missing. Considering the noisy and temporally infrequent/irregular nature of mobile phone network data, the challenge in data cleaning is still how noise/outliers can be detected and removed, without mistakenly removing the actual trip points. Methods of signal processing, such as Kalman filtering, may be employed. Map matching, which is often used in GPS data cleaning, might also be an interesting technique to be comprehensively considered. Information about the mobile network (e.g., location and spatial coverage of antennas) should also be considered during the data cleaning process. At the same time, reference datasets (e.g., GPS trajectories or detailed travel surveys) should be collected concurrently to evaluate the performance of the data cleaning methods proposed.
- **Mode detection algorithms:** Existing mode detection algorithms are mostly developed to identify easy-to-detect modes, or more general mode groups (e.g., rail versus road, moving versus stationary). Rule-based methods and unsupervised clustering methods are often employed. The challenge here is in developing algorithms that are able to differentiate more modes, especially those presenting similar speed profiles and following a same network. Supervised machine learning methods (both generative and discriminative) and fuzzy logic based methods, as shown to be effective for GPS-based mode detection, might be worth investigating (but require large amount of accurately labeled ground truth data). Methods applied in activity detection from smartphone sensor data or mobile phone network data might also give some hints on this issue, especially on the use of land use data or other auxiliary data (e.g., other sensor data than GPS, traditional travel surveys) to improve the detection results (Diao et al., 2016; Widhalm et al., 2015). Furthermore, considering that many applications making use of transport mode detection (e.g., traffic monitoring, disaster management, and event management) require real-time or near real-time results, it is also important to develop real-time mode detection algorithms, making use of the streaming location data.
- **“Benchmark” datasets and framework/metrics for evaluation:** Due to the lack of ground truth data, that is, trips labelled with their modes, most existing studies either did not report the evaluation/validation, or simply validated their proposed methods with aggregated data such as mode share statistics or farecard data. Validation with aggregated data provides only a relatively general impression on the method performance, and may be used as a “partial” evaluation when individually labeled datasets are not available. However, it is important to note that it cannot provide insights on how well the proposed methods are able to detect transport modes of individual trips. To advance the state-of-the-art of transport mode detection with mobile phone network data, it is thus vitally important to develop a collection of publicly available “benchmark” datasets, allowing different researchers to evaluate and compare their methods. This is very challenging due to the need of ensuring the “benchmark” datasets covering different transport modes and contexts, the resulting magnitude of data collection, and the legal efforts behind making such datasets available (particularly regarding the preservation of privacy). However, this may be possible, considering that many mobile network operators are now very interested in exploring the multiple usage of mobile phone network data in response to diminishing returns in their traditional business areas. Meanwhile, in addition to “benchmark” datasets, it is also important to develop “benchmark” evaluation frameworks and performance metrics that are robust and generally accepted by the scientific community.

- **Privacy:** Mobile phone network data often consist of sensitive location information, which raises privacy concerns. Currently, to address this issue, mobile phone network data are often anonymized before the data analysis process. However, as shown in de Montjoye et al. (2013), knowing four spatio-temporal points allows to uniquely re-identify a particular user with 95% probability. Therefore, sophisticated techniques that are compatible with regulations or laws (e.g., the General Data Protection Regulation in the EU) should be developed to protect the privacy of individuals. There are several initial studies addressing this issue (Blondel et al., 2015), such as re-generating random identifiers at regular time intervals (e.g., every 24 hours), k-anonymity, and location obfuscation. However, since these existing methods tend to downgrade the data quality, it is important to investigate how they impact transport mode detection.
- **Biases of the data:** Some key biases can be identified for mobile phone network data, mainly due to the lack of sociodemographic details of the subscribers (due to privacy issues), and the uneven distribution of mobile phone ownership, of cell towers in space, and of people's phone usage in space and time. For example, mobile phone ownership and usage are skewed due to social-economic, cultural, and demographic factors. There are also many people owning more than one phone. Similarly, people might use their phones more often at certain places and during certain time periods. Some of these biases, especially the distribution of cell towers in space, and people's phone usage in space and time, should be considered when developing algorithms for transport mode detection. However, all these biases have significant impacts on the further applications making use of mode detection. This is also closely linked to the more general question regarding the significance of information extracted from mobile phone network data. Theoretical and empirical analysis of different influencing factors should be pursued to address the issues brought by these biases. To give a more comprehensive picture on people's travel behaviors to support various transportation applications, it is also important to integrate mobile phone network data together with other travel survey data, such as those from traditional travel surveys, automated fare collection systems, smartphone sensor data, and even geo-social media data (e.g., (Alexander et al., 2015)).

6. Conclusion

This paper presented an in-depth, systematic review of methods used for transport mode detection with mobile phone network data. The analysis of state-of-the-art focused on the following aspects: mobile phone data used and their characteristics, data pre-processing (data cleaning and trip identification), and mode detection (algorithms, features used, and evaluation).

From the review, we can draw the conclusion that research on transport mode detection using mobile phone network data is still at an early stage. Existing studies mainly focused on identifying easy-to-detect modes, or modes aggregated into more general groups (e.g., public versus private transport). Many of them developed simple rule-based methods for mode detection, making use of geographic data of the study area. Meanwhile, due to the lack of ground truth data, evaluation of the proposed methods was seldom done and reported.

Considering the advantages of using mobile phone network data for mobility studies as well as the improvement of the data quality that took place, and will continue to take place as a consequence of advances in mobile telecommunication technology (from 3G to 4G to 5G), we trust that research on this topic will gain significant interests in the coming years. To advance the state-of-the-art, substantial research efforts are still needed, particularly in: 1) developing robust data cleaning methods to detect and remove noise/outliers, without mistakenly removing the actual trip points; 2) developing (real-time) algorithms that are able to differentiate fine-grained modes, especially those presenting similar speed profiles and following the same network; 3) providing "benchmark"

datasets and framework/metrics to allow comparison of different mode detection methods; and 4) investigating the issues of privacy and data biases. By sufficiently addressing the above aspects, we will be able to improve transport mode detection at the population scale and thus contribute to better understanding how whole populations move around in space over time.

References

- Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology* 17, 3–27. <https://doi.org/10.1080/10630731003597306>
- Alexander, L., Jiang, S., Murga, M., González, M.C., 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, Big Data in Transportation and Traffic Engineering 58, 240–250. <https://doi.org/10.1016/j.trc.2015.02.018>
- Asgari, F., 2016. Inferring user multimodal trajectories from cellular network metadata in metropolitan areas (PhD Dissertation). Institut National des Télécommunications.
- Bantis, T., Haworth, J., 2017. Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics. *Transportation Research Part C: Emerging Technologies* 80, 286–309. <https://doi.org/10.1016/j.trc.2017.05.003>
- Bayir, M.A., Demirbas, M., Eagle, N., 2010. Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing, Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns* 6, 435–454. <https://doi.org/10.1016/j.pmcj.2010.01.003>
- Blondel, V.D., Decuyper, A., Krings, G., 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4, 10. <https://doi.org/10.1140/epjds/s13688-015-0046-0>
- Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies* 17, 285–297. <https://doi.org/10.1016/j.trc.2008.11.004>
- Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J., 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems, Special Issue: Advances in Geocomputation* 36, 526–537. <https://doi.org/10.1016/j.compenvurbsys.2012.06.001>
- Bricka, S., Bhat, C.R., 2006. Comparative Analysis of Global Positioning System–Based and Travel Survey–Based Data. *Transportation Research Record* 1972, 9–20. <https://doi.org/10.1177/0361198106197200102>
- Budgen, D., Brereton, P., 2006. Performing Systematic Literature Reviews in Software Engineering, in: *Proceedings of the 28th International Conference on Software Engineering, ICSE '06*. ACM, New York, NY, USA, pp. 1051–1052. <https://doi.org/10.1145/1134285.1134500>
- Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C., 2011a. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems* 12, 141–151. <https://doi.org/10.1109/TITS.2010.2074196>
- Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies* 26, 301–313. <https://doi.org/10.1016/j.trc.2012.09.009>
- Calabrese, F., Ferrari, L., Blondel, V., 2014. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Comput. Surv.* 47, 25:1–25:20. <https://doi.org/10.1145/2655691>
- Calabrese, F., Lorenzo, G.D., Liu, L., Ratti, C., 2011b. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing* 10, 36–44. <https://doi.org/10.1109/MPRV.2011.41>
- Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., 2014. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* 61.
- Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies* 46, 326–337. <https://doi.org/10.1016/j.trc.2014.07.001>
- Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies* 68, 285–299. <https://doi.org/10.1016/j.trc.2016.04.005>
- Chin Jiaqi, K., 2018. Transport mode detection using Cellular Signaling Data (Case study of Graz and Vienna, Austria) (Master thesis). University of Zurich, Zurich.

- Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C: Emerging Technologies* 86, 360–371. <https://doi.org/10.1016/j.trc.2017.11.021>
- Danafar, S., Piorkowski, M., Kryszczuk, K., 2017. Bayesian framework for mobility pattern discovery using mobile network events, in: 2017 25th European Signal Processing Conference (EUSIPCO). Presented at the 2017 25th European Signal Processing Conference (EUSIPCO), pp. 1070–1074. <https://doi.org/10.23919/EUSIPCO.2017.8081372>
- de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3, 1376. <https://doi.org/10.1038/srep01376>
- Diao, M., Zhu, Y., Ferreira, J., Ratti, C., 2016. Inferring individual daily activities from mobile phone traces: A Boston example. *Environ Plann B Plann Des* 43, 920–940. <https://doi.org/10.1177/0265813515600896>
- Doyle, J., Hung, P., Kelly, D., McLoone, S.F., Farrell, R., 2011. Utilising Mobile Phone Billing Records for Travel Mode Discovery. Presented at the ISSC 2011, Trinity College Dublin.
- Eftekhari, H.R., Ghatte, M., 2016. An inference engine for smartphones to preprocess data and detect stationary and transportation modes. *Transportation Research Part C: Emerging Technologies* 69, 313–327. <https://doi.org/10.1016/j.trc.2016.06.005>
- Ester, M., Kriegel, H., S, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: *Proceedings of KDD 1996*. AAAI Press, pp. 226–231.
- Feng, T., Timmermans, H.J.P., 2013. Transportation mode recognition using GPS and accelerometer data. *Transportation Research Part C: Emerging Technologies* 37, 118–130. <https://doi.org/10.1016/j.trc.2013.09.014>
- García, P., Herranz, R., Javier, J., 2016. Big Data Analytics for a Passenger-Centric Air Traffic Management System. Presented at the 6th SESAR Innovation Days, Delft, Netherlands.
- González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. *Nature* 453, 779–782. <https://doi.org/10.1038/nature06958>
- Hollecze, T., The Anh, D., Yin, S., Jin, Y., Antonatos, S., Goh, H.L., Low, S., Shi-Nash, A., 2015. Traffic Measurement and Route Recommendation System for Mass Rapid Transit (MRT), in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*. ACM, New York, NY, USA, pp. 1859–1868. <https://doi.org/10.1145/2783258.2788590>
- Horn, C., Gursch, H., Kern, R., Cik, M., 2017. QZTool—Automatically Generated Origin-Destination Matrices from Cell Phone Trajectories, in: Stanton, N.A., Landry, S., Di Buccianico, G., Vallicelli, A. (Eds.), *Advances in Human Aspects of Transportation, Advances in Intelligent Systems and Computing*. Springer International Publishing, pp. 823–833.
- Horn, C., Kern, R., 2015. Deriving Public Transportation Timetables with Large-Scale Cell Phone Data. *Procedia Computer Science, The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015)* 52, 67–74. <https://doi.org/10.1016/j.procs.2015.05.026>
- Horn, C., Klampfl, S., Cik, M., Reiter, T., 2014. Detecting Outliers in Cell Phone Data: Correcting Trajectories to Improve Traffic Modeling. *Transportation Research Record* 2405, 49–56. <https://doi.org/10.3141/2405-07>
- Hui, K.T.Y., Wang, C., Kim, A., Qiu, T.Z., 2017. Investigating the Use of Anonymous Cellular Phone Data to Determine Intercity Travel Volumes and Modes. Presented at the Transportation Research Board 96th Annual Meeting Transportation Research Board.
- Hui, T.Y., 2017. Investigating the Use of Anonymous Cellular Data for Intercity Travel Patterns in Alberta (Master thesis). University of Alberta. <https://doi.org/10.7939/R3QN5ZR65>
- Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., Smoreda, Z., 2013. Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies, in: Vandenbroucke, D., Bucher, B., Cromptoets, J. (Eds.), *Geographic Information Science at the Heart of Europe, Lecture Notes in Geoinformation and Cartography*. Springer International Publishing, Cham, pp. 247–265. https://doi.org/10.1007/978-3-319-00615-4_14
- Järv, O., Tenkanen, H., Toivonen, T., 2017. Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science* 31, 1630–1651. <https://doi.org/10.1080/13658816.2017.1287369>
- Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J., Jr., Frazzoli, E., González, M.C., 2013. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities, in: *Proceedings of the 2Nd*

- ACM SIGKDD International Workshop on Urban Computing, UrbComp '13. ACM, New York, NY, USA, pp. 2:1–2:9. <https://doi.org/10.1145/2505821.2505828>
- Kalatian, A., Shafahi, Y., 2016. Travel Mode Detection Exploiting Cellular Network Data. MATEC Web Conf. 81, 03008. <https://doi.org/10.1051/mateconf/20168103008>
- Larijani, A.N., Olteanu-Raimond, A.-M., Perret, J., Brédif, M., Ziemlicki, C., 2015. Investigating the Mobile Phone Data to Estimate the Origin Destination Flow and Analysis; Case Study: Paris Region. Transportation Research Procedia, 4th International Symposium of Transport Simulation (ISTS'14) Selected Proceedings, Ajaccio, France, 1-4 June 2014 6, 64–78. <https://doi.org/10.1016/j.trpro.2015.03.006>
- Lee, J.-K., Hou, J.C., 2006. Modeling Steady-state and Transient Behaviors of User Mobility: Formulation, Analysis, and Application, in: Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc '06. ACM, New York, NY, USA, pp. 85–96. <https://doi.org/10.1145/1132905.1132915>
- Li, G., Chen, C.-J., Peng, W.-C., Yi, C.-W., 2017. Estimating Crowd Flow and Crowd Density from Cellular Data for Mass Rapid Transit, in: Proceedings of The 6th International Workshop on Urban Computing (in Conjunction with ACM KDD 2017).
- Louail, T., Lenormand, M., Cantu Ros, O.G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2014. From mobile phone data to the spatial structure of cities. Scientific Reports 4, 5276. <https://doi.org/10.1038/srep05276>
- Miao, G., Zander, J., Sung, K.W., Slimane, S.B., 2016. Fundamentals of Mobile Data Networks. Cambridge University Press.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, T.P., 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine 6, e1000097. <https://doi.org/10.1371/journal.pmed.1000097>
- Naboulsi, D., Fiore, M., Ribot, S., Stanica, R., 2016. Large-Scale Mobile Traffic Analysis: A Survey. IEEE Communications Surveys Tutorials 18, 124–161. <https://doi.org/10.1109/COMST.2015.2491361>
- Newson, P., Krumm, J., 2009. Hidden Markov Map Matching Through Noise and Sparseness, in: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '09. ACM, New York, NY, USA, pp. 336–343. <https://doi.org/10.1145/1653771.1653818>
- Ogle, J., Guensler, R., Elango, V., 2005. Georgia's Commute Atlanta Value Pricing Program: Recruitment Methods and Travel Diary Response Rates. Transportation Research Record: Journal of the Transportation Research Board 1931, 28–37. <https://doi.org/10.3141/1931-04>
- Oliver, N., Matic, A., Frias-Martinez, E., 2015. Mobile Network Data for Public Health: Opportunities and Challenges. Front Public Health 3. <https://doi.org/10.3389/fpubh.2015.00189>
- Petticrew, M., Roberts, H., 2006. Systematic Reviews in the Social Sciences: A Practical Guide. Blackwell Pub.
- Phithakkitnukoon, S., Sukhvibul, T., Demissie, M., Smoreda, Z., Natwichai, J., Bento, C., 2017. Inferring social influence in transport mode choice using mobile phone data. EPJ Data Science 6, 11. <https://doi.org/10.1140/epjds/s13688-017-0108-6>
- Poonawala, H., Kolar, V., Blandin, S., Wynter, L., Sahu, S., 2016. Singapore in Motion: Insights on Public Transport Service Level Through Farecard and Mobile Data Analytics, in: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. ACM, New York, NY, USA, pp. 589–598. <https://doi.org/10.1145/2939672.2939723>
- Prelipcean, A.C., Gidófalvi, G., Susilo, Y.O., 2017. Transportation mode detection – an in-depth review of applicability and reliability. Transport Reviews 37, 442–464. <https://doi.org/10.1080/01441647.2016.1246489>
- Prelipcean, A.C., Gidofalvi, G., Susilo, Y.O., 2016. Measures of transport mode segmentation of trajectories. International Journal of Geographical Information Science 30, 1763–1784. <https://doi.org/10.1080/13658816.2015.1137297>
- Qu, Y., Gong, H., Wang, P., 2015. Transportation Mode Split with Mobile Phone Data, in: 2015 IEEE 18th International Conference on Intelligent Transportation Systems. Presented at the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 285–289. <https://doi.org/10.1109/ITSC.2015.56>
- Rojas, M.B., Sadeghvaziri, E., Jin, X., 2016. Comprehensive Review of Travel Behavior and Mobility Pattern Studies That Used Mobile Phone Data. Transportation Research Record: Journal of the Transportation Research Board 2563, 71–79. <https://doi.org/10.3141/2563-11>
- Schlaich, J., Otterstatter, T., Friedrich, M., 2010. Generating Trajectories from Mobile Phone Data. Presented at the Transportation Research Board 89th Annual MeetingTransportation Research Board.

- Schönfelder, S., Axhausen, K.W., Antille, N., Bierlaire, M., 2002. Exploring the potentials of automatically collected GPS data for travel behaviour analysis: a Swedish data source. Institut für Geoinformatik, Universität Münster. <https://doi.org/10.3929/ethz-a-004403386>
- Schuessler, N., Axhausen, K., 2009. Processing Raw Data from Global Positioning Systems Without Additional Information. *Transportation Research Record: Journal of the Transportation Research Board* 2105, 28–36. <https://doi.org/10.3141/2105-04>
- Shad, S.A., Chen, E., Bao, T., 2012. Cell Oscillation Resolution in Mobility Profile Building. *International Journal of Computer Science Issues* 9, 205–213.
- Shen, L., Stopher, P.R., 2014. Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews* 34, 316–334. <https://doi.org/10.1080/01441647.2014.903530>
- Smoreda, Z., Olteanu Raimond, A.-M., Couronné, T., 2013. Spatiotemporal data from mobile phones for personal mobility assessment, in: Zmud, J., Lee-Gosselin, M., Carrasco, J.A., Munizaga, M.A. (Eds.), *Transport Survey Methods: Best Practice for Decision Making*. Emerald Group Publishing Limited, pp. 745–767.
- Steenbruggen, J., Borzacchiello, M.T., Nijkamp, P., Scholten, H., 2013. Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. *GeoJournal* 78, 223–243. <https://doi.org/10.1007/s10708-011-9413-y>
- Wang, F., Chen, C., 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies* 87, 58–74. <https://doi.org/10.1016/j.trc.2017.12.003>
- Wang, H., Calabrese, F., Lorenzo, G.D., Ratti, C., 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records, in: 13th International IEEE Conference on Intelligent Transportation Systems. Presented at the 13th International IEEE Conference on Intelligent Transportation Systems, pp. 318–323. <https://doi.org/10.1109/ITSC.2010.5625188>
- Wang, Z., He, S.Y., Leung, Y., 2018. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society* 11, 141–155. <https://doi.org/10.1016/j.tbs.2017.02.005>
- Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C., 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42, 597–623. <https://doi.org/10.1007/s11116-015-9598-x>
- Wolf, J., 2000. Using GPS data loggers to replace travel diaries in the collection of travel data (PhD Dissertation). Georgia Institute of Technology.
- Wu, L., Yang, B., Jing, P., Wu, L., Yang, B., Jing, P., 2016. Travel Mode Detection Based on GPS Raw Data Collected by Smartphones: A Systematic Review of the Existing Methodologies. *Information* 7, 67. <https://doi.org/10.3390/info7040067>
- Wu, W., Cheu, E.Y., Feng, Y., Le, D.N., Yap, G.E., Li, X., 2013. Studying intercity travels and traffic using cellular network data. Presented at the Mobile Phone Data for Development: Net Mob 2013.
- Wu, W., Wang, Y., Gomes, J.B., Anh, D.T., Antonatos, S., Xue, M., Yang, P., Yap, G.E., Li, X., Krishnaswamy, S., Decraene, J., Nash, A.S., 2014. Oscillation Resolution for Mobile Phone Cellular Tower Data to Enable Mobility Modelling, in: 2014 IEEE 15th International Conference on Mobile Data Management. Presented at the 2014 IEEE 15th International Conference on Mobile Data Management, pp. 321–328. <https://doi.org/10.1109/MDM.2014.46>
- Xu, D., Song, G., Gao, P., Cao, R., Nie, X., Xie, K., 2011. Transportation Modes Identification from Mobile Phone Data Using Probabilistic Models, in: Tang, J., King, I., Chen, L., Wang, J. (Eds.), *Advanced Data Mining and Applications, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 359–371.
- Yamada, Y., Uchiyama, A., Hiromori, A., Yamaguchi, H., Higashino, T., 2016. Travel estimation using Control Signal Records in cellular networks and geographical information, in: 2016 9th IFIP Wireless and Mobile Networking Conference (WMNC). Presented at the 2016 9th IFIP Wireless and Mobile Networking Conference (WMNC), pp. 138–144. <https://doi.org/10.1109/WMNC.2016.7543981>
- Yuan, Y., Raubal, M., 2016. Exploring Georeferenced Mobile Phone Datasets – A Survey and Reference Framework. *Geography Compass* 10, 239–252. <https://doi.org/10.1111/gec3.12269>
- Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., Yin, L., 2016. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science* 30, 1738–1762. <https://doi.org/10.1080/13658816.2015.1137298>